orrick faegre drinker

April 3, 2025

VIA ECF

Honorable Ona T. Wang
United States District Court
500 Pearl Street
New York, NY 10007

Re: *Authors Guild v. OpenAI Inc.*, 23-cv-8292 (S.D.N.Y.), and *Alter v. OpenAI Inc.*, 23-cv-10211 (S.D.N.Y.): Response to Brief Regarding Request to Inspect Late 2023 Data (ECF 377)

Dear Magistrate Judge Wang:

Microsoft opposes Class Plaintiffs' letter motion to compel for inspection purported "training data [Microsoft] provided to OpenAI in late 2023." The Court should deny the motion for two reasons: (1) Microsoft does not possess or control copies of the data that Plaintiffs seek to inspect; and (2) even if Microsoft had a copy, the data is not relevant to Plaintiffs' claims.

I. Plaintiffs' Request Has Been A Moving Target.

When Plaintiffs initiated discussions regarding the data inspection request, it was entirely unclear to Microsoft what information Plaintiffs specifically sought to inspect. Plaintiffs first raised the inspection request in mid-February, premised upon unspecified documents creating the "appear[ance] that Microsoft provided web crawling data to OpenAI and that the crawler data captured books datasets." *See* Ex. A (February 13 email from Plaintiffs). They did not identify any RFPs providing a basis for the informal request and did not serve a request for inspection. Microsoft asked for further explanation, and Plaintiffs sent a copy of an October 2023 document (now marked as Exhibit 1 to their motion) that [redacted]. *See* Ex. B (February 21 email from Plaintiffs); ECF 377-1. In response, Microsoft explained that the preliminary, exploratory activities discussed in the document are irrelevant to this Action, as they occurred well after the conduct alleged in the Complaint and referred to data that was not used for training the GPT models at issue in this case. ECF 377-9 at 2, 4. In reply, Plaintiffs identified three documents "predating 2023 to provide further context for the request," all of which were plainly unrelated to the document they originally identified as forming the basis of the request. *See* ECF 377-10 at 5; ECF 377-3 ([redacted]); ECF 377-5 at 3 ([redacted]); Ex. C (MSFT_AICPY_000003514) ([redacted]).

Only now in their letter brief do Plaintiffs state that they seek to inspect "the training data [Microsoft] provided to OpenAI ***in late 2023***." Plaintiffs newly rely on Exhibits 2, 4, and 6, which were not provided to Microsoft during any of the parties' conferrals. Now that Plaintiffs have clearly identified what they are seeking, it is apparent that the motion should be denied because

Microsoft does not possess or control the material at issue and in all events it is irrelevant to Plaintiffs' claims.

II. Microsoft Does Not Possess or Control the Data Plaintiffs Seek to Inspect.

Plaintiffs seek inspection of [redacted]. *See* Ex. D. *Id.* Microsoft does not possess or control copies of [redacted] As Microsoft does not possess or control the material collected, the Court should deny Plaintiffs' motion. *Jackson v. Edwards*, No. 99CIV.0982(JSR)(HBP), 2000 WL 782947, at *3-4 (S.D.N.Y. June 16, 2000).

III. The Data Plaintiffs Seek to Compel for Inspection Are Irrelevant to Plaintiffs' Allegations Set Forth in the Complaint.

Additionally, even if Microsoft possessed copies of the scraped internet data that Plaintiffs seek to inspect, the collection is wholly irrelevant to Plaintiffs' claims. Simply put, data collected in late 2023 were not used to train any of the models at issue in this case (GPT-3, GPT-3.5, GPT-4, and GPT-4 Turbo). Rather, the late 2023 data were, at most, potentially related to OpenAI's plans for training future models. RFP 41, upon which Plaintiffs rely, only seeks documents that were actually used to train LLMs at issue.

Plaintiffs do not even contest that they seek discovery into models untethered to their allegations. *See* ECF 377-6 at 4 [redacted]) (emphasis added). The only basis for entitlement to this data Plaintiffs set forth in support of their motion is that, as of late 2023, Microsoft "may have" provided data to OpenAI that OpenAI [redacted]. Plaintiffs then stretch this to [redacted]. This chained series of inferences provide no reasonable basis to compel inspection. ECF 377 at 1. The relevance of this data is pure speculation, thus providing no basis for a finding that inspection would be proportional to the needs of the case. Plaintiffs' request is at best remotely connected to future models that were not identified in the current Complaint. The Court has denied similar discovery sought from Microsoft by the News Plaintiffs.

Plaintiffs' attempt to use the discovery letter briefing process to expand the scope of the case should likewise be denied. First, Plaintiffs mischaracterize a prior Order on a discovery dispute between Plaintiffs and OpenAI, construing it as a ruling that the GPT models within the scope of the case extend beyond those specifically identified in the Complaint. The court did not so rule. Indeed, rather than ruling on the relevance of "later models like GPT-4o . . . and GPT-4o Mini," the Court denied Plaintiffs' demand for discovery into any models that OpenAI did not identify in its interrogatory response and made no such ruling as to the relevance of any models. *See* ECF 293 at 2. Although OpenAI agreed to discovery into these later models, OpenAI has not conceded their relevance. *See* ECF 281 at 1 ("This lawsuit, filed by book authors, centers around specific books datasets that were used to train GPT-3 and 3.5. OpenAI agreed to provide discovery into these and a number of *other* models—including GPT-1, GPT-2, GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo—and, beyond that, to supplement its response to Interrogatory 11 . . . .") (cleaned up). But even if it had, any such concession would not be binding on Microsoft—Plaintiffs assert

different claims against Microsoft that correspond to a different scope of relevance and different burdens.

Plaintiffs also contend that there is some "Microsoft copy of LibGen" that somehow falls within the direct and contributory infringement conduct alleged. As a threshold matter, Plaintiffs' direct infringement claim is based on the allegation of reproducing Plaintiffs' works "***in datasets used to train [Defendants'] artificial intelligence models***." *See* ECF 69, ¶ 415 (emphasis added). Plaintiffs now seem to argue that they "are entitled to inspect the data because [redacted] constitutes an act of direct infringement" on its own, irrespective of any model training. *See* ECF 377 at 3. Whether or not there is such a viable claim, Plaintiffs cannot use discovery in this to investigate it nor can they add it to this case via a discovery motion.

Indeed, on March 19, Class Plaintiffs provided to Microsoft a proposed Amended Complaint that would dramatically expand the scope of the case to include models trained by Microsoft Research as well as new products released long after the lawsuit was filed. Microsoft declined to stipulate to the filing of the amended pleading because it would, conservatively estimated, triple the size of the case and derail the schedule and make the case unmanageable. Class Plaintiffs did not file their motion for leave to amend. And even under the proposed amendment that didn't happen, the data collected in 2023 for OpenAI had nothing to do with the proposed additions as to Microsoft.

Nor can Plaintiffs rely on the contributory infringement claims as a springboard into irrelevant discovery. To reiterate, each and every claim asserted in the Complaint is cabined to conduct relating to the relevant OpenAI models. The notion that the mere "existence and contents of such a dataset"—created years after OpenAI completed training the relevant GPT models—evinces Microsoft's knowledge that OpenAI was using particular datasets for training years earlier is nonsensical. As Plaintiffs admit, liability for contributory infringement can be had only for "persons who 'know or have reason to know ***of the direct infringement***.'" ECF 377 at 3 (quoting *Abbey House Media, Inc. v. Apple Inc.*, 66 F. Supp. 3d 413, 419 (S.D.N.Y. 2014)) (emphasis added). Plaintiffs do not, and cannot, tie the "existence and contents" of late 2023 data to any act of alleged infringement at issue in this case.

Plaintiffs are not entitled to boundless discovery into subject matter that is entirely unrelated the claims they assert. The motion should be denied.

Respectfully submitted,	Respectfully submitted,
/s/ Annette L. Hurst	*/s/ Jared B. Briant*
Annette L. Hurst	Jared B. Briant

Counsel for Defendant Microsoft Corporation